BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 14 2010, pages 1797–1799
doi:10.1093/bioinformatics/btq278

Systems biology

ABC-SysBio—approximate Bayesian computation in Python with GPU support

Juliane Liepe[1,†], Chris Barnes[1,2,∗,†], Erika Cule[1,3,†], Kamil Erguler[1], Paul Kirk[1,2], Tina Toni[1,2,∗] and Michael P.H. Stumpf[1,2,4,∗]

[1]Centre for Bioinformatics, Division of Molecular Biosciences, [2]Institute of Mathematical Sciences, [3]Department of Epidemiology and Public Health, School of Public Health and [4]Centre for Integrative Systems Biology, Imperial College London, London, UK.

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The growing field of systems biology has driven demand for flexible tools to model and simulate biological systems. Two established problems in the modeling of biological processes are model selection and the estimation of associated parameters. A number of statistical approaches, both frequentist and Bayesian, have been proposed to answer these questions.

Results: Here we present a Python package, ABC-SysBio, that implements parameter inference and model selection for dynamical systems in an approximate Bayesian computation (ABC) framework. ABC-SysBio combines three algorithms: ABC rejection sampler, ABC SMC for parameter inference and ABC SMC for model selection. It is designed to work with models written in Systems Biology Markup Language (SBML). Deterministic and stochastic models can be analyzed in ABC-SysBio.

Availability: http://abc-sysbio.sourceforge.net

Contact: christopher.barnes@imperial.ac.uk; ttoni@imperial.ac.uk; m.stumpf@imperial.ac.uk

Received on February 5, 2010; revised on April 16, 2010; accepted on May 24, 2010

1 INTRODUCTION

In the last decade, modeling of biochemical systems using ordinary and stochastic differential equations (ODE and SDE) has become increasingly popular as quantitative ideas have begun to pervade the biomolecular sciences. Inferring model parameters and ranking alternative models is necessary in order to gather reliable future predictions about the dynamical behavior of such systems. The problem of parameter estimation in deterministic systems has been addressed by using local and global non-linear optimization methods (Mendes and Kell, 1998; Moles *et al*., 2003) as well as maximum-likelihood estimation (Baker *et al*., 2005; Bortz and Nelson, 2006; Muller *et al*., 2004; Timmer and Muller, 2004) and within a Bayesian framework (Banks *et al*., 2005, Huang *et al*., 2006, Putter *et al*., 2002). An approximate Bayesian computation (ABC) scheme based on sequential Monte Carlo (SMC) has been developed for likelihood-free parameter inference in deterministic and stochastic systems (Toni *et al*., 2009). Furthermore, this approach also allows for model selection, i.e. evidence-based ranking of alternative models.

Because in ABC methods the evaluation of the likelihood is replaced by simulations, the implementation of these methods hides some numerical and technical problems. This includes finding numerically the solution of stiff ODE systems or stochastic systems described by SDEs or a master equation. This requires a flexible and adaptive implementation of ABC methods to address each specific model with its biochemical as well as dynamical problems.

Here we present an extensible Python package, ABC-SysBio, which implements approximate Bayesian computation for parameter inference and model selection in deterministic and stochastic models. The package supports the standard models exchange format, SBML, as well as user-defined models written in Python. In addition, graphical processing unit support is provided via pycuda (Klöckner *et al*., 2009). User-defined algorithmic parameters allow for the adaptation and tuning of the inference procedures included in the package to suit each specific dynamical model.

Note that in contrast to other packages for parameter inference in a Bayesian framework, e.g. DIYABC (Cornuet *et al*., 2008) and BioBayes (Vyshemirsky and Girolami, 2008), ABC-SysBio allows for parameter inference and model selection of both deterministic and stochastic models. Its implementation is flexible and user friendly: it supports important model exchange standards and is easily integrated into existing computational pipelines and systems biology frameworks through the flexibility of Python.

2 METHODS

The package ABC-SysBio is implemented as a Python module, abcsysbio. Together with the two Python scripts abc-sysbio-sbml-sum and run-abc-sysbio, it creates a user friendly tool that can be applied to models in SBML format without any further software development. It is advisable to use the package together with the Python Enthought Distribution, though this is not essential. It works on both MacOS and Linux operating systems.

The module, abcsysbio, can be imported into an interactive Python session, and by defining the arguments to the functions in the interactive namespace, they can be used through the Python shell.

When run-abc-sysbio is called, model(s) written in SBML format are parsed to generate a corresponding Python module representing the model. The format of the Python module written depends on the integration type, which also informs the program which solver to use to simulate the model.

We provide algorithms to simulate ODE, SDE and Gillespie models (Press *et al*., 1992). All algorithms are adapted to the specific requirements of

∗To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First authors.

models in the Biomodels database. Beside the possibility of only numerically solving the provided model, one of the three following algorithms can be called.

2.1 ABC rejection sampler for parameter inference

Given a parameter θ, its prior distribution $\pi(\theta)$ and a dataset x, we want to approximate the posterior distribution $\pi(\theta|x)$. The ABC rejection sampler proceeds as follows:

(1) Sample θ^* from π.

(2) Simulate a dataset x^* from the model with parameter θ^*.

(3) If $d(x_0,x^*) \leq \epsilon$, accept θ, otherwise reject.

(4) Return to 1.

where $d(x_0,x^*)$ is a distance function and ϵ is a tolerance. The implemented distance function is the Euclidian distance; however the user can easily define custom distance metrics. To obtain reliable parameter estimates, ϵ should be very small. The ABC rejection sampler should be used only for simple systems that allow a fast simulation, because the rejection rate of this algorithm is usually very high.

The ABC rejection sampler describes the first sampled population of the ABC SMC algorithm and is therefore implemented as part of the ABC SMC algorithms described below.

2.2 ABC SMC for parameter inference

Using ABC in a SMC framework leads to the ABC SMC algorithm (Toni *et al*., 2009). A number of particles are sampled from the prior distribution $\pi(\theta)$ and propagated through a sequence of intermediate distributions until the population represents an approximated posterior distribution. The intermediate distributions are defined by a sequence of tolerances ϵ_t in decreasing order, $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_T \geq 0$. Therefore ABC SMC for parameter inference is automatically invoked when only one model, but more than one ϵ are provided. Additionally, the user needs to provide perturbation kernels, which define how a particle is perturbed after resampling from the previous intermediate distribution.

This algorithm can be considered a special case of the model selection algorithm implemented in ABC-SysBio. The ABC SMC for parameter inference algorithm is nested within ABC SMC for model selection. Therefore, the same ABC-SysBio functions, with the same computational features, are called whichever of the algorithms are run.

2.3 ABC SMC for model selection

An algorithm to select between several deterministic or stochastic dynamical models for a given dataset has been implemented; here the model identifiers/labels are treated as an additional parameter, as described in Toni and Stumpf, 2010. Therefore, a prior distribution over models as well as perturbation kernels for model transitions need to be defined by the user. The implementation of the model selection algorithm using ABC SMC provides the framework for the above mentioned nested algorithms.

2.4 Options

Users can define several algorithmic parameters, for example, prior distributions or perturbation kernels. Several distributions are already implemented, e.g. uniform distribution, Gaussian distribution and log-normal distribution, but further distributions can be added easily. The user should define the ϵ schedule, because it is strongly dependent on the biochemical/dynamical system under investigation, as well as on the noise in the provided data. Note that the dataset does not necessarily need to include data for all species defined in the model, but can be a subset or even a combination of several species.

After each sampled population, ABC-SysBio provides the user with information about the algorithm's progress. The rejection rate per population,

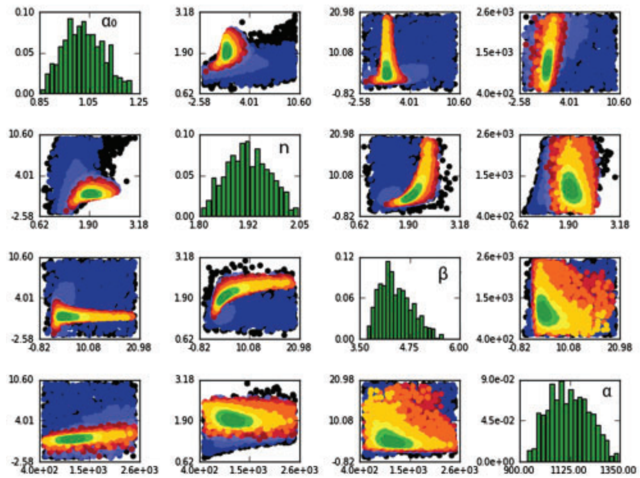

Fig. 1. ABC-SysBio output. This example shows the results of the deterministic repressilator model (as described in Toni *et al*., 2009). Scatterplots of inferred parameters for 11 populations are displayed (from black for the first population to dark green for the 9th population). The approximate posterior over model parameters is shown on the diagonal.

as well as the sampled particles from intermediate distributions (Fig. 1) are saved in accessible text files. Furthermore, a graphical output including scatterplots of pairwise parameter combinations and histograms summarizing parameter and model distributions are provided. This allows the user to follow the progress of the algorithm. Additionally, all data are copied into a binary file to allow the algorithm to be restarted from a previous population, with changed algorithm parameters.

3 SUMMARY

ABC-SysBio is a flexible, extendable and user-friendly Python package that can import models described in the SBML standard combined with experimental data. Our package approximates the posterior parameter distribution and compares different model structures to select the model that represents the data best.

ACKNOWLEDGEMENTS

We thank the members of the *Theoretical Systems Biology Group* at Imperial College London, Sylvia Richardson, David Balding, Mark Beaumont, Christian Robert and Scott Sisson for discussions on ABC methodology. We are particularly grateful to Justina Norkunaite for becoming an early adopter of ABC-SysBio.

Funding: Wellcome Trust (to J.L., E.C., P.K. and K.E.); Biotechnology and Biological Sciences Research Council (to C.B. and M.P.H.S.); Medical Research Council (to T.T.). M.P.H.S. is a Royal Society Wolfson Research Merit Award holder.

Conflict of Interest: none declared.

REFERENCES

Baker,C. *et al*. (2005) Ludewig computational approaches to parameter estimation and model selection in immunology. *J. Comput. Appl. Math.*, **184**, 5076.

Banks,H. *et al*. (2005) A hierarchical Bayesian approach for parameter estimation in HIV models. *Inverse Probl.*, **21**, 1803–1822.

Bortz,D.M. and Nelson,P.W. (2006) Model selection and mixed-effects modeling of HIV infection dynamics. *Bull. Math. Biol.*, **68**, 2005–2025.

Cornuet,J.-M. *et al*. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Huang,Y. *et al*. (2006) Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, **62**, 413–423.

Klöckner,A. *et al*. (2009) PyCUDA: GPU run-time code generation for high-performance computing. Available at: http://arxiv.org/abs/0911.3456 (last accessed date November 18, 2009).

Mendes,P. and Kell,D. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, **14**, 869–883.

Moles,C. *et al*. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.*, **13**, 2467–2474.

Muller,T.G. *et al*. (2004) Tests for cycling in a signalling pathway. *J. R. Stat. Soc. Ser. C*, **53**, 557.

Press,W.H. *et al*. (1992) Numerical Recipes in C: The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge.

Putter,H. *et al*. (2002) A Bayesian approach to parameter estimation in HIV dynamical models. *Stat. Med.*, **21**, 2199–2214.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Timmer,J. and Muller,T. (2004) Modeling the nonlinear dynamics of cellular signal transduction. *Int. J. Bifurcat. Chaos*, **14**, 2069–2079.

Toni,T. *et al*. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

Toni,T. and Stumpf,MPH. (2010) Simulation-based model selection for dynamical systems and population biology. *Bioinformatics*, **26**, 104–110.

Vyshemirsky,V. and Girolami,M. (2008) Biobayes: a software package for bayesian inference in systems biology. *Bioinformatics*, **24**, 338–1934.